Breast Cancer Prediction Using Soft Voting Classifier Based on Machine Learning Models

Mohammed S. Hashim, Ali A.Yassin

Abstract— **Breast cancer is one of the most serious illnesses that many individuals worldwide face. Accurate detection and effective treatment are of vital significance in lowering the death rate of breast cancer. Although researchers throughout the globe have offered many diagnostic approaches for the identification of this illness, these current methods still need additional refinement to assure the proper and efficient diagnosis of this disease. The purpose of this study is to make early and precise forecasts regarding breast cancer, which consider the second biggest cause of death in women globally. which in turn reduces the number of deaths around the world. In this study, we propose a methodology that utilises a soft voting classifier for diagnosing the type of breast cancer tumor, whether benign or malignant, based on three machine learning algorithms, namely, logistic regression, support vector machine, and decision tree. The Wisconsin Diagnostic Breast Cancer dataset was used to assess the proposed methodology. Before using this dataset, we balanced it using the SMOTE technique to eliminate bias and increase the size of the dataset. Modern studies have been surpassed by the proposed methodology, which has 99.3% accuracy, 100% precision, 98.46% recall, 99.2% F1 score, and an AUC of 0.992. Furthermore, it achieved an accuracy mean of 97.24% with 10-fold cross-validation.**

Index Terms— **Breast Cancer, Predication, Soft Voting, Cross-validation**

I. Introduction

Artificial intelligence (AI) in healthcare is an umbrella term used to describe the application of cognitive technologies in medical settings. In the simplest sense, AI is when computers and other machines mimic human cognition and become capable of learning, thinking and making decisions or taking actions. In particular, AI has significant applications in diagnostics and prediction. AI can help physicians and medical providers to detect and diagnose accurately a disease and establish treatment plans depending on patient's information. As a result, AI-based healthcare is more predictive and proactive because it analyses big data to develop improved preventive care recommendations for patients.

Breast cancer is a serious and severe illnesses. According to the report of the World Health Organisation (WHO), 2.3 million women was diagnosed with breast cancer in 2020, and 685,000 related deaths were recorded worldwide. In addition, the number of new cases of breast cancer diagnosed will increase by 70% over the next 20 years[1]. Breast cancer ranks as the fifth most lethal disease following lung, colon, liver and stomach cancers. Female breast cancer (FBC), which will account for 2.3 million new cancer cases (11.7% of all cases), is the most common cancer, according to the Global Cancer Statistics 2020 (GLOBOCAN)[2].

Benign (limited tumour growth at a single location) and malignant (the tumour moves to other body parts and damages the healthy tissues) cancers can be distinguished. Breast cancer is due to the aberrant development of breast cells. Several techniques have been developed for the precise diagnosis of breast cancer. Mammography and breast screening can be used to identify breast cancer[3]. Women’s nipple status can also be assessed using X-ray. Early-stage breast cancer has fairly unnoticeable external signs, making the diagnosis challenging. A simple test that can identify cancer at an early stage to use in the mammography.

Breast cancer has no known treatment. Removal of the damaged body part is the only method to save the life of a person with breast cancer. In this regard, the best methods and mechanisms must be used for the early detection of breast cancer to expedite the removal of the tumour before it spreads[3].

Machine learning is a method that has been very helpful in several areas, including the prediction of early-stage breast cancer. Early detection will significantly increase the percentage of survival because the chances of surviving vary dramatically by stages of breast cancer. The chances of survival are higher for women who received a diagnosis earlier than for those who receive one later.

Several studies (details in the related works section) were conducted to diagnose whether a tumour is benign or malignant by using the Wisconsin Diagnostic Breast Cancer dataset. However, these researches suffer from the accuracy limitations of machine learning classifiers because they used an unbalanced (original) dataset that biased the machine learning models towards the majority category [4]. Thus, models will be constrained by their ability to predict minority classes. Another constraint is the use of each model separately and the comparison of the results of these models. Therefore, a mechanism that combines the advantages of these models should be used to obtain the best diagnosis.

This study's primary contributions are mentioned as follows: First, we used SMOTE as a way to balance the dataset. This helps get rid of bias that happens during training, which improves the accuracy of classification. Secondly, we proposed the soft voting classifier as a

Manuscript received October 11, 2022; revised February 22, 2023.

M. S. Hashim is a Postgraduate Student of Department of Computer Science, Education College for Pure Sciences, University of Basrah, Basrah, 61004 Iraq (moh.salah@uobasrah.edu.iq).

A. A. Yassin is a professor of Department of Computer Science, Education College for Pure Sciences, University of Basrah, Basrah, 61004 Iraq (Ali.Yassin@uobasrah.edu.iq).

classification model, which includes three models, as these are considered the best three models to work with according to a set of experiments. This classifier integrates these three models into a single model that carries the power of these models, which in turn improves classification accuracy. This research aims mainly to reduce death rates and identify tumours in their early stages. Fig. 1 shows the differences between the traditional and our proposed architecture.

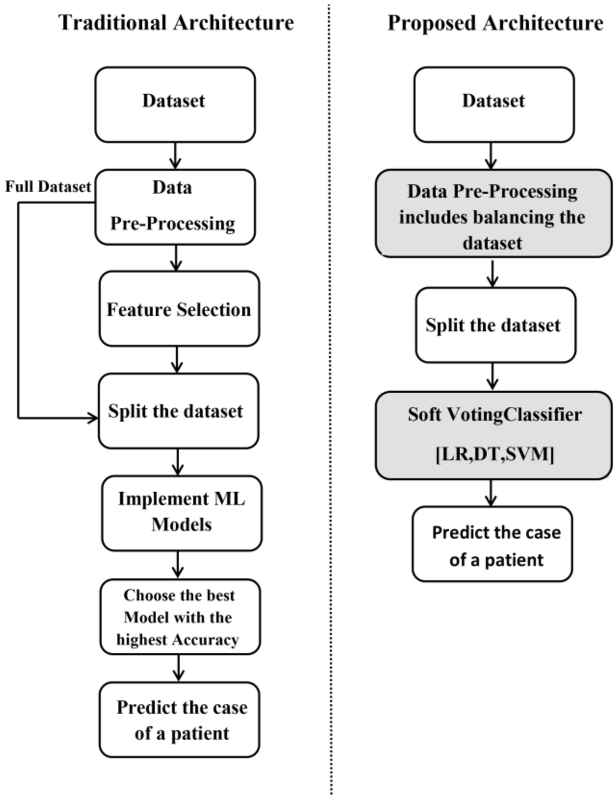

Traditional Architecture
Dataset
Full Dataset
Data Pre-Processing
Feature Selection
Split the dataset
Implement ML Models
Choose the best Model with the highest Accuracy
Predict the case of a patient
Proposed Architecture
Dataset
Data Pre-Processing includes balancing the dataset
Split the dataset
Soft VotingClassifier [LR,DT,SVM]
Predict the case of a patient

Fig. 1. Differences between the traditional and our proposed architecture

The paper's remaining sections are organised as follows: Section II explains the machine learning models and evaluation metrics used in the study. Section III presents the related works of previous studies on the diagnosis of breast cancer. Section IV explains the proposed methodology. Section V shows the evaluation of the results and discussions. Section VI provides the conclusion.

II. Background

This section explains the machine learning (ML) models and evaluation metrics used in the study.

A. Prediction Models

A prediction model is the basis of every prediction process, and many machine learning models have been used to predict and diagnose breast cancer. In this section, we will discuss the models used, namely, logistic regression, decision tree, and support vector machine.

Logistic Regression

Logistic regression (LR) is a statistical model that depicts the relationship between an independent variable and a qualitative dependent variable that can only take certain discrete values. LR models are utilised to explore the effect of predictor variables on categorical outcomes. When the outcome is binary, such as the presence or absence of an illness, the model is referred to as a binary logistic model [5]. In this model, a series of explanatory factors are linked to the likelihood of a level. This model is used to analyse datasets with one or more independent factors that affect the outcome. Given a collection of independent variables, the model is used to forecast a binary result (in which the possible outcomes are two). Fig. 2 shows the mechanism of LR using the sigmoid function.

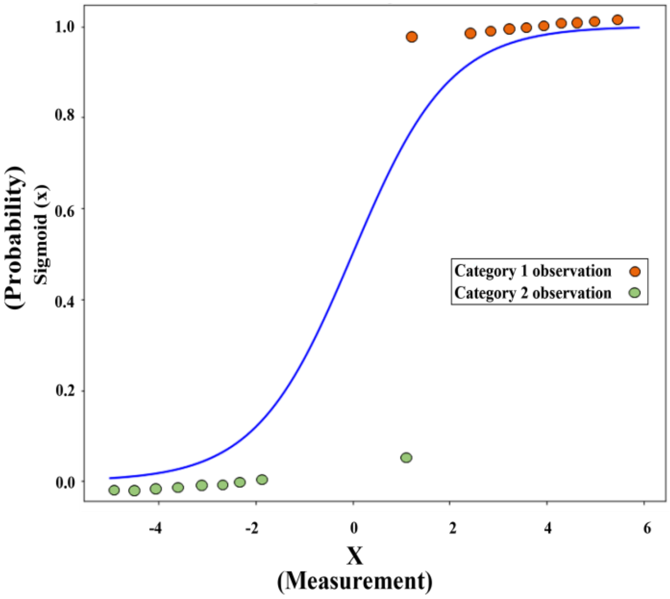


Fig. 2. Representation of logistic regression algorithm[6]

Decision Tree

A decision tree (DT) is one of the supervised learning algorithms that are widely used in machine learning. DT is mainly used in two-field classification and regression. It solves classification problems by drawing a tree from top to bottom. A DT consists of a root node, a decision node and a leaf node. Each decision node represents a feature, and the leaf nodes represent the output. Decision nodes are arranged from top to bottom based on certain criteria to determine the best decision node**.** Information gain and entropy are calculated to determine the best feature. From each level of DT, we select the attribute with the highest information gain to be the current decision node, as shown in the following equations [7]**:**

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i) \quad (1)$$

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times I(D_j) \quad (2)$$

$$Gain(A) = Info(D) - Info_A(D) \quad (3)$$

Support Vector Machine

Support vector machine (SVM), which belongs to supervised learning algorithms, is one of the most powerful machine learning algorithms. SVM aims to build a linear separator between two data points to distinguish two different classes in a multi-dimensional environment[8]. It draws a line between the two categories known as a linear classifier. SVM defines the margin of a linear classifier as the width that the boundary can be increased by before hitting a data point. The maximum margin linear classifier is the linear classifier of SVM [9]. SVM has many kernels; in this work, we use the RBF kernel. Fig. 3 explains the mechanism of SVM.

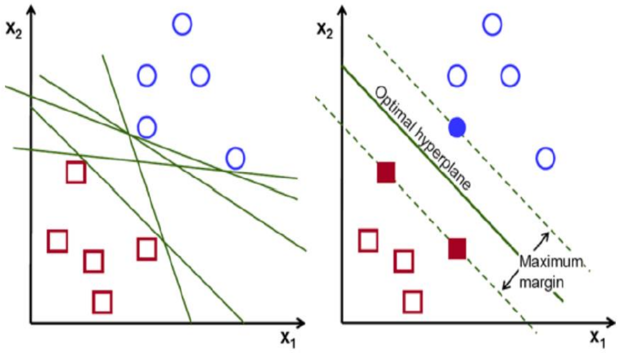


Fig. 3. Mechanism of SVM [10]

Voting Classifier

One of the ensemble techniques used to build a powerful classifier with higher classification accuracy than traditional ML classifiers is the voting classifier (VC). Ensemble-based algorithms often perform better than others on most of the datasets. In order to create a strong model that carries the power of the input models, the VC takes many artificial intelligence models and votes among their prediction outcomes. There are two forms of voting: hard and soft [11]. The voting classifier's workflow is depicted in Fig. 4.

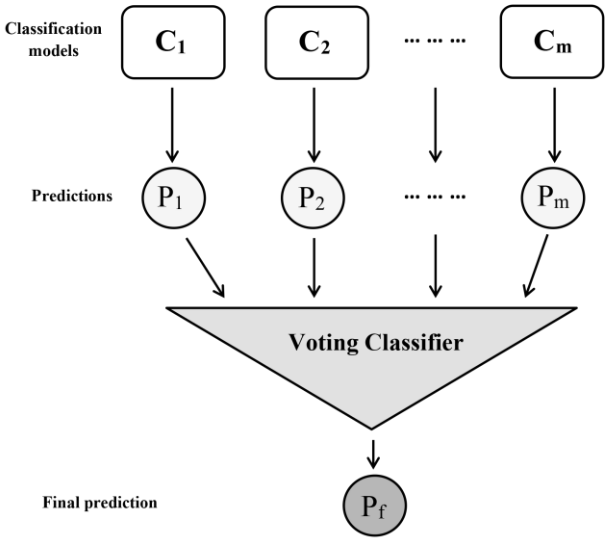


Fig. 4. Voting classifier

B. Evaluation Metrics

Building the best classifier requires careful consideration of evaluation metrics. Therefore, choosing appropriate evaluation metrics is a crucial step in making distinguishing and locating the best classifier. In this work, we use confusion matrix, recall, precision, F1 score, and accuracy to evaluate our models.

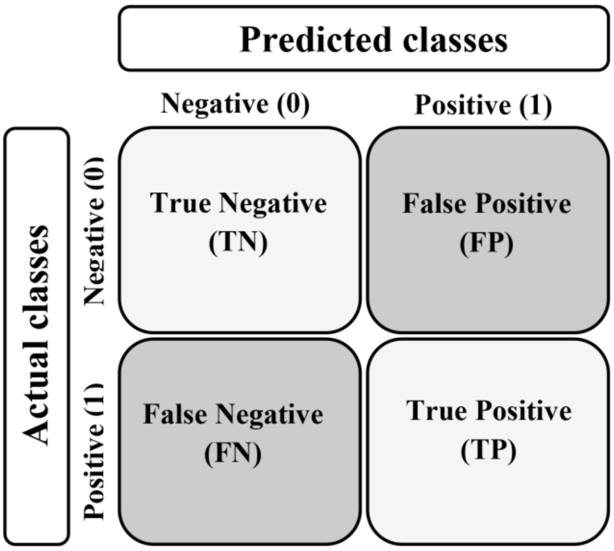
The confusion matrix is a performance measure for machine learning classification situations where the output might be two or more classes. It is a table with four separate sets of actual and predicted values [TP, FP, FN, TN], as shown in Fig. 5. The process of calculating other measures depends on the results of the values; therefore, the values of the measures (recall, precision, F1 score, and accuracy] can be calculated using the following equations [12]**:**

$$\text{Recall} = \text{TP}/(\text{TP}+\text{FN}) \tag{4}$$

$$\text{Precision} = \text{TP}/(\text{TP}+\text{FP}) \tag{5}$$

$$\text{F1 Score} = 2*[(\text{Precision}*\text{Recall})/(\text{Precision}+\text{Recall})] \tag{6}$$

$$\text{Accuracy} = [(\text{TP}+\text{TN})/(\text{TP}+\text{TN}+\text{FP}+\text{FN})] \tag{7}$$



Actual classes	Predicted classes: Negative (0)	Predicted classes: Positive (1)
Negative (0)	True Negative (TN)	False Positive (FP)
Positive (1)	False Negative (FN)	True Positive (TP)

Fig. 5. Confusion matrix

III. Related Works

The healthcare field is one of the most important areas in which AI has been applied due to the urgent need for accuracy in diagnosis. Many experiments were conducted on breast cancer datasets using machine learning and deep learning algorithms and obtained accurate classification.

Naji et al.[13] performed a study on the WDBC dataset obtained from the repository (UCI) and compared the results of DT (C4.5), random forest (RF), SVM, logistic regression (LR) and K-NNs. SVM displayed the highest classification accuracy of 97.2%.

Fatih[14] compared the suggested applications for ML and data visualisation in the detection and diagnosis of breast cancer. The WDBC dataset was subjected to the implementation of naive Bayes, SVM, K-NNs, RF, DT and LR classification algorithms, and the LR model had the highest classification accuracy (98.1%). Milon et al.[15] compared the performance of five ML algorithms, namely, LR, K-NNs, RF, SVM and ANNs, on the WBC dataset. The accuracy, precision and F1 score of ANNs were the highest, with values of 98.57%, 97.82% and 0.9890, respectively.

Noor[16] evaluated the effectiveness of ML algorithms, namely, RF, SVM and multilayer perception (MLP). The WBCD Centre provided the information. Accuracy, precision and recall were used to evaluate and compare the performance of the models. MLP performed best in terms of accuracy (95.96%), precision (95.21%) and recall (96.31%).

Teixeira et al.[17] developed five distinct categorisation techniques for evaluation: deep neural network, MLP, DT and RF. They employed a database from the University of Wisconsin Hospital; it contains 30 parameters that describe the characteristics of the breast mass' nucleus. The DNN classifier performed the best in terms of accuracy (92%).

Khourdifi et al.[18] compared K-NNs, naive Bayes, RF and SVM and determined the most effective machine learning technique. The experimental results showed that SVM had the highest accuracy of 97.9%. Sakri et al.[19] used the Wisconsin Breast Cancer Prognostic Dataset to develop and compare five phase-based data analytical approaches. RepTree, NB and K-NNs obtained accuracy of 76.3%, 70.0% and 66.3%, respectively. They also utilised Weka as a tool for data analysis. Banu and Subramanian[20] highlighted the use of naive Bayes methods to predict the presence of breast cancer; they compared tree augmented naive Bayes (TAN), Bayes belief network (BBN) and boosted augmented naive Bayes (BAN). Statistical Analytical Software Enterprise Miner (SAS-EM) was used to implement the models on the well-known WDBC dataset. TAN, BAN and BBN achieved accuracy of 94.11%, 91.7% and 91.7%, respectively, with the aid of gradient boosting. Thus, TAN was found to be the most accurate classifier for this dataset among naive Bayes approaches. Chaurasia et al.[21] utilized the WBC dataset and three well-known algorithms, namely, RBF Network, J48 and naive Bayes, to create a prediction model. The holdout sample findings showed that naive Bayes is the top predictor with the highest accuracy of 97.36%, followed by RBF Network (96.77%) and then J48 (93.41%).

Aruna et al.[22] utilised DTs, SVM and naive Bayes on the WDBC dataset; they reported that SVM had the best results, with an accuracy of 98.06%.

Asri et al. [23] tested K-NNs, naive Bayes, DT (C4.5), and SVM on the same dataset (WBC) and compared their performance in terms of accuracy, specificity, precision and sensitivity. The experimental findings indicated that SVM had the highest accuracy of 97.13% and the best score.

The following are the main contributions of our paper:

1) Our work focuses on balancing the dataset to build a good, bias-free model; this ensures that machine learning models are learned correctly, which helps improve diagnostic accuracy.
2) We proposed the soft voting classifier as a classification model, which includes three models, as these are considered the best three models to work with according to a set of experiments. This classifier integrates these three models into a single model that carries the power of these models, which in turn improves classification accuracy.
3) A good model is proposed by looking at the prognostic features of patients with early-stage breast cancer from a wider perspective and comparing the models' strengths by using accurate measurements.
4) Data visualisation and machine learning technologies are used to diagnose and detect breast cancer, but a more thorough comparison and analysis are required to validate the model.

IV. Methodology

In this section, we show the four main phases of the proposed architecture, namely, pre-processing phase, split phase, training phase and prediction phase, to build a suitable model for detection of breast cancer with high accuracy and help physicians to follow up patients as shown in Fig. 6.

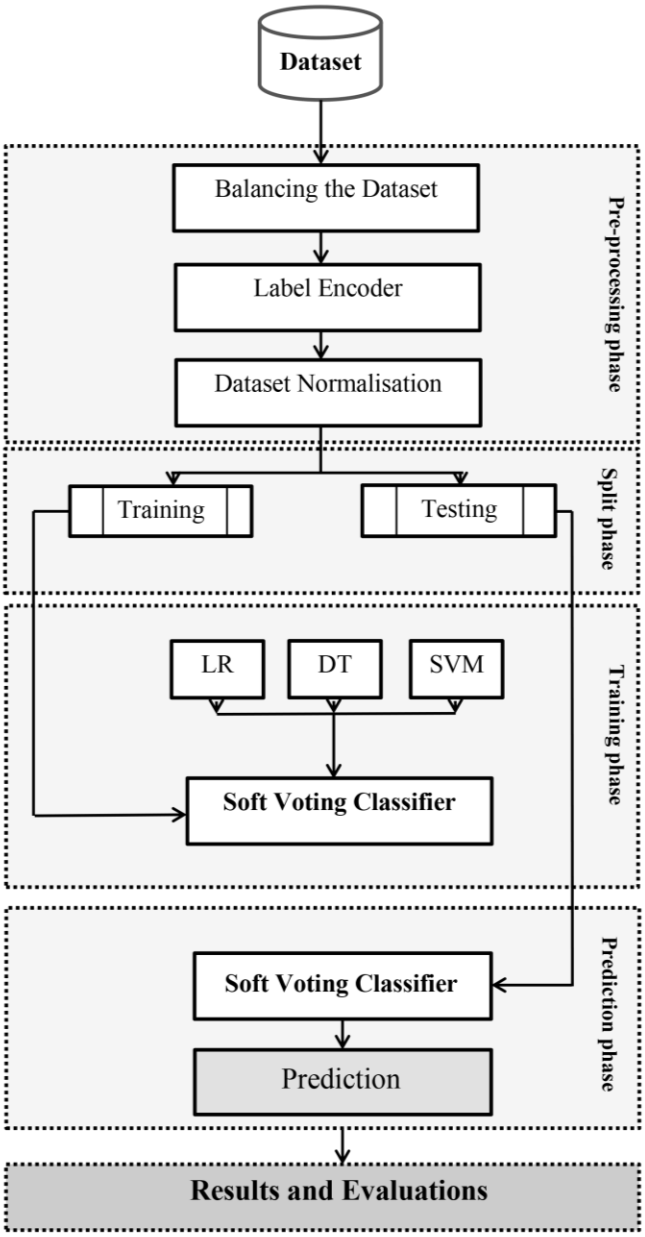


Fig. 6. Proposed architecture

A. Dataset Description

We used the Wisconsin Diagnostic Breast Cancer dataset (WDBC) from the University of Wisconsin Hospital's

Madison Breast Cancer Database[24]. The dataset consists of 32 columns and 569 examples divided into two classes: malignant (212 instances, 37.26%) and benign (357 instances, 62.74%). Table I presents the main features of the dataset as follows:

TABLE I
FEATURES OF THE WDBC DATASET

Feature_Name	**Type**	**Feature_Name**	**Type**
id	Number	smoothness_se	Number
diagnosis	Category	compactness_se	Number
radius_mean	Number	concavity_se	Number
texture_mean	Number	concave points_se	Number
perimeter_mean	Number	symmetry_se	Number
area_mean	Number	fractal_dimension_se	Number
smoothness_mean	Number	radius_worst	Number
compactness_mean	Number	texture_worst	Number
concavity_mean	Number	perimeter_worst	Number
concave points_mean	Number	area_worst	Number
symmetry _ mean	Number	smoothness_worst	Number
fractal _dimension _ mean	Number	compactness_worst	Number
radius_se	Number	concavity _worst	Number
texture _ se	Number	concave points _ worst	Number
perimeter _ se	Number	symmetry_worst	Number
area_se	Number	fractal_dimension_worst	Number

The above features were obtained from the digitised images of breast mass obtained through FNA. The values of each feature represent the characteristics and shape of the cell nucleus so we can determine and diagnose breast cancer.

B. Pre-processing phase

This phase improves the dataset quality so valuable insights may be drawn. We organise the raw data to create and train the machine learning models. The procedures carried out at this stage are listed below.

Balancing the Dataset

An imbalanced dataset causes machine learning models to be biased towards the majority class. In this regard, we used SMOTE to improve the performance of the models. In this technique, samples in the feature space that are close to one another are chosen, a line is drawn between them and a new sample is then drawn at a location on the line. Fig. 7 shows the mechanism of SMOTE [25].

The main goal of this approach is to balance the dataset and eliminate the disparity between the majority and minority classes, thereby removing skew in favour of the majority class in the classifier training and the most common metric for evaluating classification quality.

The original dataset consists of 569 cases divided into two classes, namely, 212 malignant cases and 357 benign cases. After using SMOTE, the balanced dataset consists of 714 cases divided into two classes, namely, 357 malignant (M) cases and 357 benign (B) cases as depicted in Fig. 8.

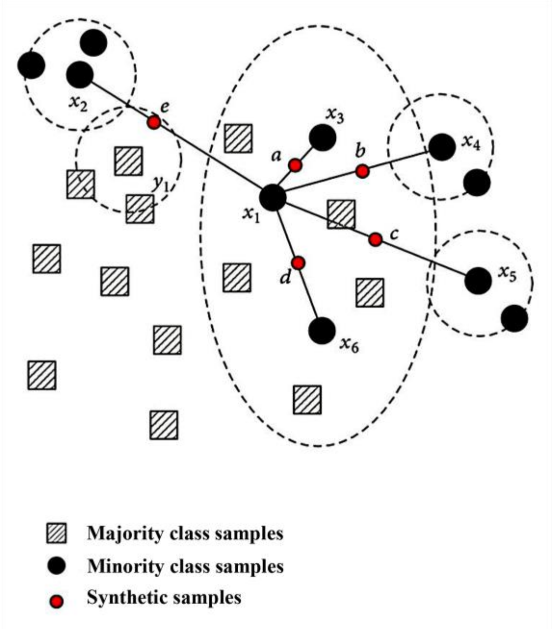
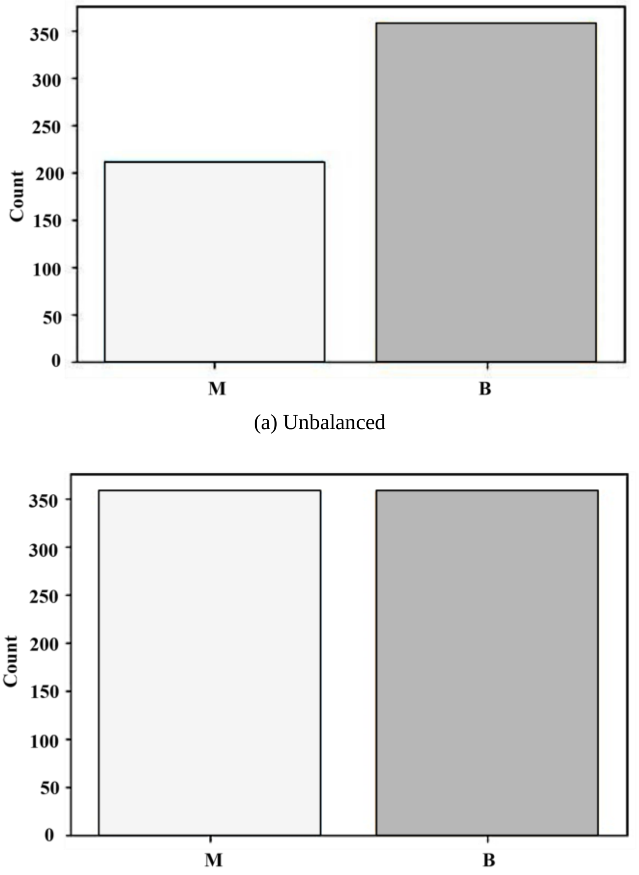


Fig. 7. SMOTE[25]



(a) Unbalanced

(b) Balanced

Fig. 8. Dataset before and after using SMOTE

Label Encoder

ML models require numerical input and output variables. As such, we encode the label (diagnosis) through the dataset balancing procedure using the Label Encoder function. Categorical data should be encoded into integers before the training and assessment of a model. As illustrated in Fig. 9, we will change each benign case into 0 and each malignant case into 1.

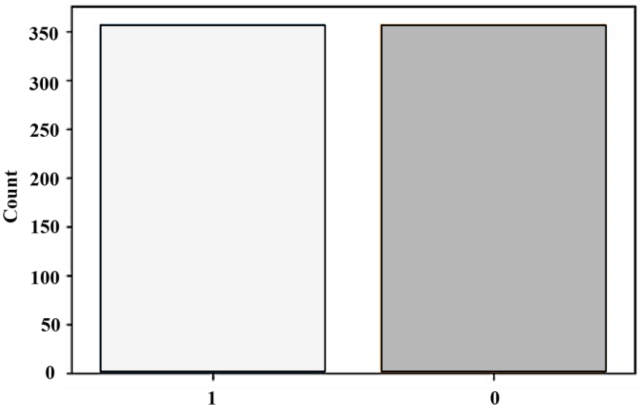
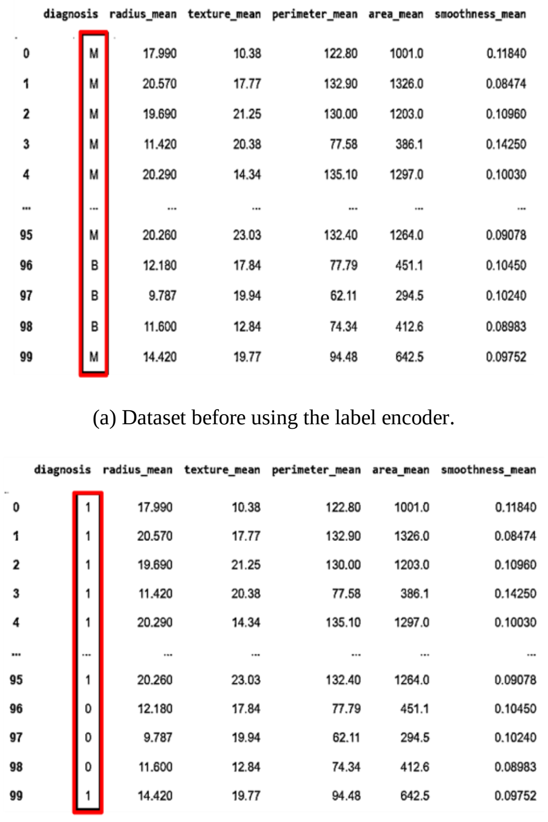


Fig. 9. Label encoder

Fig. 10 shows a part of the dataset before and after applying the label encoder.



	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean
0	M	17.990	10.38	122.80	1001.0	0.11840
1	M	20.570	17.77	132.90	1326.0	0.08474
2	M	19.690	21.25	130.00	1203.0	0.10960
3	M	11.420	20.38	77.58	386.1	0.14250
4	M	20.290	14.34	135.10	1297.0	0.10030
...	...	...	...	...	...	...
95	M	20.260	23.03	132.40	1264.0	0.09078
96	B	12.180	17.84	77.79	451.1	0.10450
97	B	9.787	19.94	62.11	294.5	0.10240
98	B	11.600	12.84	74.34	412.6	0.08983
99	M	14.420	19.77	94.48	642.5	0.09752

(a) Dataset before using the label encoder.

	diagnosis	radius_mean	texture_mean	perimeter_mean	area_mean	smoothness_mean
0	1	17.990	10.38	122.80	1001.0	0.11840
1	1	20.570	17.77	132.90	1326.0	0.08474
2	1	19.690	21.25	130.00	1203.0	0.10960
3	1	11.420	20.38	77.58	386.1	0.14250
4	1	20.290	14.34	135.10	1297.0	0.10030
...	...	...	...	...	...	...
95	1	20.260	23.03	132.40	1264.0	0.09078
96	0	12.180	17.84	77.79	451.1	0.10450
97	0	9.787	19.94	62.11	294.5	0.10240
98	0	11.600	12.84	74.34	412.6	0.08983
99	1	14.420	19.77	94.48	642.5	0.09752

(b) Dataset after using the label encoder.

Fig. 10. Effect of the label encoder on a dataset

Dataset Normalisation

Outliers may sometimes appear in the dataset. Machine learning models benefit from the standardisation of the dataset. Individual features may behave poorly if the dataset does not resemble standard normally distributed data; in this regard, we employed **StandardScaler** to scale the features. Scaling the features is an essential step in modelling the algorithms with the datasets because the features have several dimensions and scales. The modelling of a dataset is hampered by the different scales of the data components. Prediction outcomes are skewed as a function of misclassification error and accuracy. Therefore, scaling the data is necessary before modelling. As shown in (8), the primary goal is to scale to unit variance and eliminate the mean[26].

$$z = d/s \tag{8}$$

Where:

- $d = x - u$
- x: is the sample
- u: is the samples' mean
- s: is the samples' standard deviation

C. Split Phase

We firstly divide the dataset into x (which represents all features without the target) and y (which represents the target). We will then split the dataset into two parts, namely, training and testing sets, by using the **Train _Test _ Split** Procedure. Training data are used to train the model, and testing data are used to test the performance of the model after training. Fig. 11 shows the procedure of **Train _Test _ Split** in our architecture.

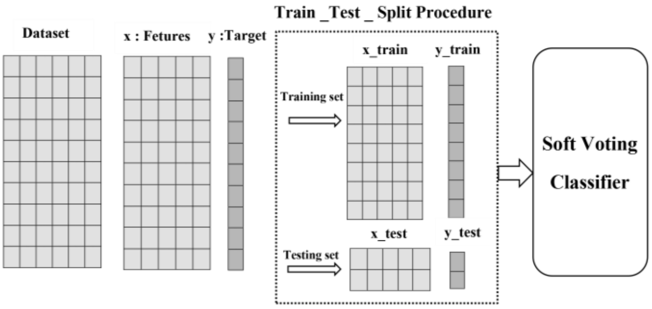


Fig. 11. Split phase

This phase is important for preparing our dataset for the next phases. After training the models using the training set, they can detect breast cancer.

D. Training Phase

The goal of this phase is to build a single model that learns from numerous models and predicts outcomes based on the overall majority of votes for each output class, as opposed to creating unique, specialized models and evaluating their accuracy. This model is known as the voting classifier model. In our work, we will use a soft voting classifier. After performing experiments, we conclude that the best three models that can pass to the voting classifier are SVM, DT, and LR, which are trained using the training set by using [Soft Voting Classifier. fit (x_train, y_train)]. After completing the training process, the model is ready for performance testing. Fig. 12 shows the main steps of the training phase which involves initially including the three selected models in the soft voting classifier and then training this classifier so that we have one model that carries the power of the three models that are included in it. Finally, the testing part is ready to check any new examples for detecting breast cancer based on the prediction phase.

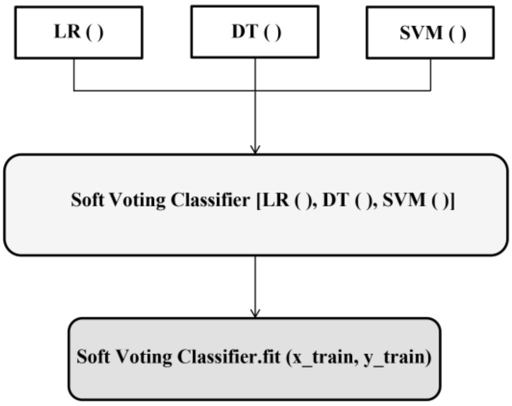

LR ( )
DT ( )
SVM ( )
Soft Voting Classifier [LR ( ), DT ( ), SVM ( )]
Soft Voting Classifier.fit (x_train, y_train)

Fig. 12. Training phase

E. Prediction Phase

After training the model, the soft voting classifier model will be ready for testing the performance. In our work, we will use the soft voting classifier that includes three models to predict class labels (**diagnosis**) based on expected probability **p**. Firstly, the soft voting classifier calculates the probability value of the class label for both classes (0 and 1) for all models. Secondly, it finds the average probabilities of the class label for both classes (0 and 1) from all models. Finally, it calculates the final prediction **(y')** by considering the maximum probability average of classes (0, 1) as the correct prediction as in (9) [27]. Fig. 13 shows the workflow.

$$y' = \arg\max_{i} \sum_{j=1}^{m} w_j \cdot p_{ij} \qquad (9)$$

- w_j is the weight that the j^{th} classifier model was given.
- p_{ij} is the probability value of both classes for all models.

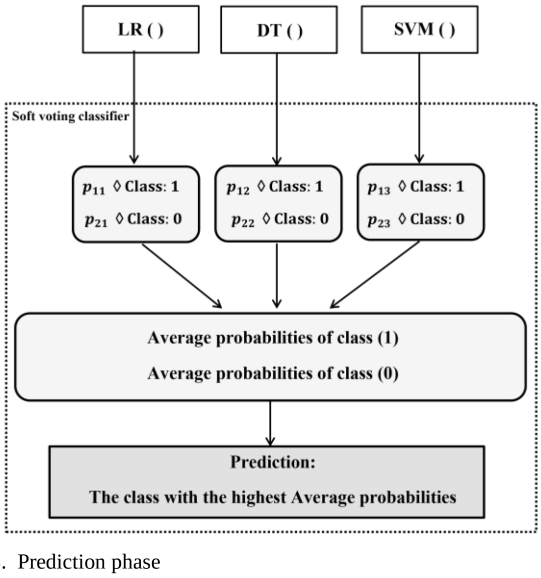


Fig. 13. Prediction phase

The soft voting classifier combines the strengths of the models [LR, DT and SVM] by taking the highest probability average to obtain a high classification accuracy.

V. Results and Discussion

In this section, we show and discuss the accuracy, F1 score, precision, recall, AUC, and ROC curve performance results we got from using the proposed methodology, where we evaluate the results of our proposed model (Soft Voting Classifier) with the results of the models included in it by using balanced datasets. Also, the results of our proposed model are evaluated on the balanced dataset through 10-fold cross-validation.

We use the balanced dataset, which contains 714 cases (M = 357, B = 357). We split the balanced dataset into two parts: testing (20%) and training (80%). After that, we train and test the performance of the soft voting classifier as well as the models that we use in this study, LR, DT, and SVM. The comparison of the results is presented in Table II.

TABLE II
COMPARISON BETWEEN OUR MODELS USING THE BALANCED DATASET

Model	Accuracy (%)	Precision	Recall	F1 Score	AUC
LR	97.90%	0.984	0.969	0.976	0.978
DT	95.80%	0.927	0.984	0.955	0.960
SVM	97.20%	0.969	0.969	0.969	0.978
Soft Voting Classifier	**99.30%**	**1**	**0.984**	**0.992**	**0.992**

Through the above table, we notice that the soft voting classifier obtained the highest values in terms of recall, precision, accuracy, F1 score, and AUC, as shown in Fig. 14, outperforming the rest of the models as their performance was tested on 143 cases. The reason for this is due to the working mechanism of the voting classifier, which collects the features and strengths of the models included in it and thus obtained the highest classification accuracy. In addition, balancing the data set has a significant impact on increasing and improving the classification accuracy, because it eliminates any bias that occurs during the training of the models, therefor it’s helped the models be properly trained.

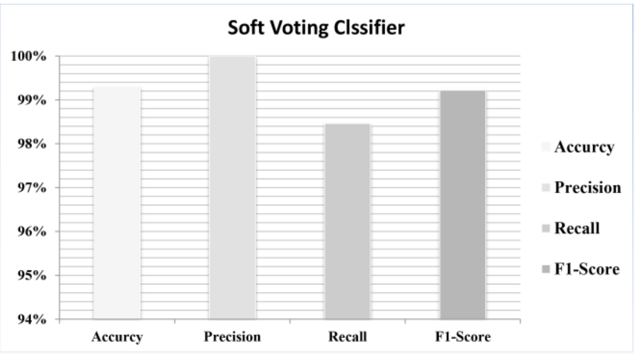


Fig. 14. Performance of soft voting classifier

Fig. 15 shows how well the soft voting classifier worked in the balanced dataset compared to other models that were used.

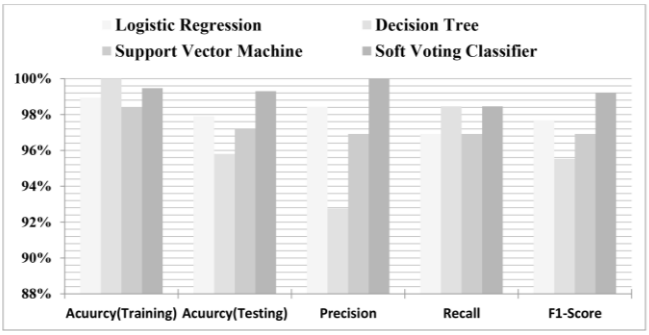
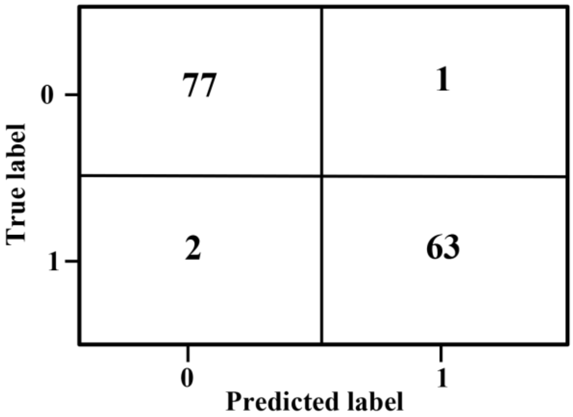


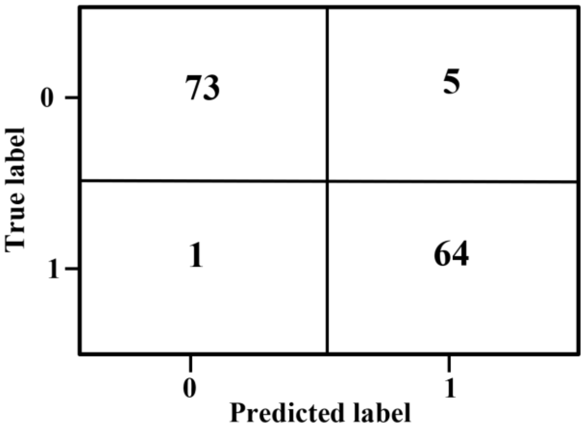
Fig. 15. Performance comparison of the models using the balanced dataset

The confusion matrix of LR, DT, SVM, and soft voting classifier shows in Fig.16, Fig.17, Fig.18, and Fig.19 respectively.



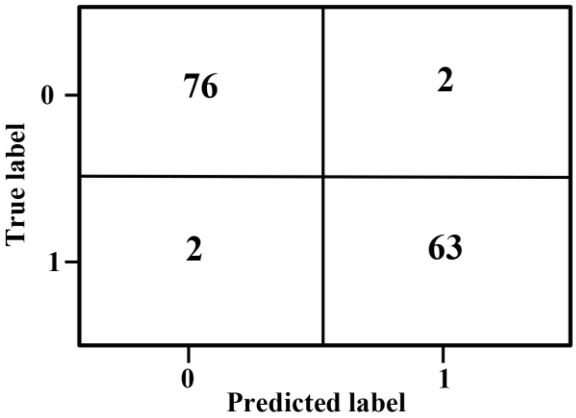
True label \ Predicted label	0	1
0	77	1
1	2	63

Fig. 16. Confusion_matrix of LR



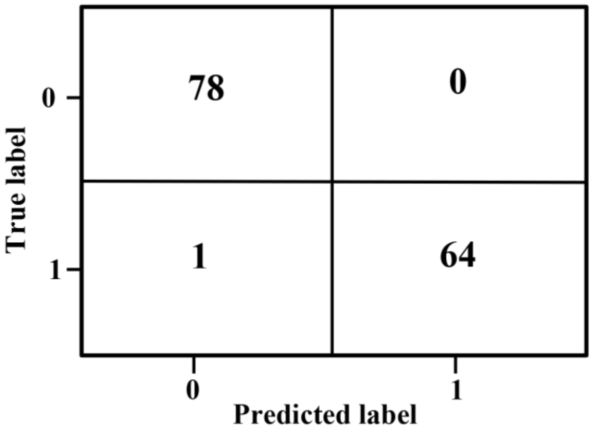
True label \ Predicted label	0	1
0	73	5
1	1	64

Fig. 17. Confusion_matrix of DT



True label \ Predicted label	0	1
0	76	2
1	2	63

Fig. 18. Confusion_matrix of SVM



True label \ Predicted label	0	1
0	78	0
1	1	64

Fig. 19. Confusion_matrix of soft voting classifier

Through the above confusion matrixes, we note that the performance of the models was tested on 143 cases, and we found the soft voting classifier made a mistake in classifying only one case, and this is evidence of the efficiency of this model compared to other models.

Fig. 20 depicts the ROC curves that examine the predictive capacity of a classifier and gives a visual method to see how changes in thresholds impact the performance of our models. And it also provides the AUC value of each model, which is used as a summary of the ROC curve and is regarded as a measure of the classifier's ability to discriminate between classes.

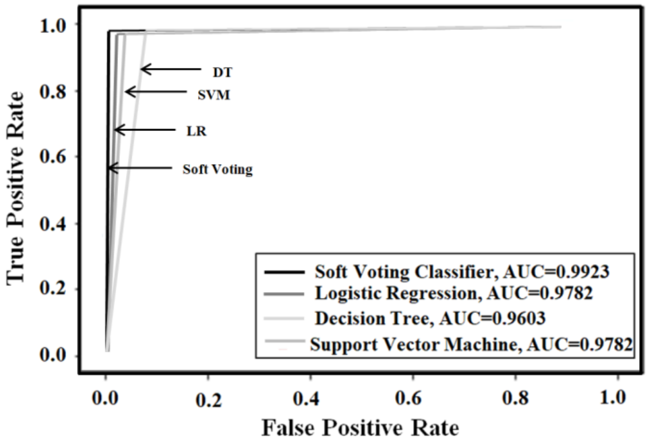


Fig. 20. AUC and ROC curves for our models

Through the above figure, we notice that the soft voting classifier obtained an AUC value of 0.9923, a logistic regression AUC value of 0.9782, a decision tree AUC value of 0.9603, and the support vector machine AUC value of 0.9782. We can tell from this that our proposed model (the soft voting classifier), which got the highest AUC value, is better than the other models.

TABLE III shows the comparison of the results of our proposed model (**Soft Voting Classifier**) and previous models based on the **WDBC** dataset.

TABLE III
COMPARISON WITH STATE-OF-THE-ART MODELS ON THE WDBC DATASET

Author	Year	Balanced the dataset	Methods	Accuracy (%)	Precision	Recall	F1 Score
Aruna[22]	2011	NO	**SVM**	**98.06%**	**-**	**-**	**-**
			Naïve Bayes	92.61%	-	-	-
			RBF networks	93.67%	-	-	-
			Decision Tree (J48)	92.97%	-	-	-
			Decision Tree(CART)	92.97%	-	-	-
Banu[20]	2018	NO	**Naïve Bayes(Tree Augmented)**	**94.11%**	**-**	**-**	**-**
			Bayes Belief Network	91.7%	-	-	-
			Naïve Bayes(Boosted Augmented)	91.7%	-	-	-
Khourdifi[18]	2018	NO	**SVM**	**97.9%**	**0.979**	**0.979**	**0.979**
			KNN	96.1%	0.961	0.961	0.961
			Random Forest	96%	0.960	0.960	0.960
			Naïve Bayes	92.6%	0.926	0.926	0.926
Fatih [14]	2020	NO	**Logistic Regression**	**98.06%**	**-**	**-**	**-**
			KNN	96.49%	-	-	-
			SVM	96.49%	-	-	-
			Naïve Bayes	94.73%	-	-	-
			Decision Tree	95.61%	-	-	-
			Random Forest	95.61%	-	-	-
			Rotation Forest	95.61%	-	-	-
Naji [13]	2021	NO	**SVM**	**97.2%**	**0.98**	**-**	**0.96**
			Random Forest	96.5%	0.96	-	0.95
			Logistic Regression	95.8%	0.98	-	0.94
			Decision tree (C4.5)	95.1%	0.94	-	0.93
			KNN	93.7%	0.92	-	0.91
Our work	**2023**	Yes	**Soft Voting Classifier [LR ,DT ,SVM]**	**99.3%**	**1**	**0.984**	**0.992**

Finally, we use the 10-fold cross-validation technique to evaluate our model for estimating the out-of-sample error. In each round, we use 10–1 fold for training and the remaining fold for testing. The results were calculated by taking the mean of the model scores. The k-fold cross-validation helps avoid overfitting and builds a generalised model to better evaluate the performance of our model. This technique is applied to the proposed "soft voting classifier" and the balanced data set. The results showed a mean accuracy of 97.24% in testing.

VI. Conclusion

Our study proposes a model based on the voting strategy used (LR, DT and SVM) for accurate, efficient and early prediction of breast cancer. Our work adds balance to the original dataset (WDBC). The proposed model outperformed other state-of-the-art models when implemented in the same dataset, with accuracy of 99.3%, precision of 100%, recall of 98.46%, F1 score of 99.2%, and AUC of 0.992. A 10-fold cross-validation comparison was conducted, where the proposed model had accuracy of 97.24%, which is higher than those of other reported models.

Our research focuses on a simple sample of the population; hence, the results cannot be extrapolated to a larger population. Future research should focus on clinical datasets, predictions, models and feature selection techniques.

REFERENCES

[1] World Health Orgnaization | WHO, “http://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/,” *World Breast Cancer Rep.*, 2020.

[2] H. Sung *et al.*, “Global Cancer Statistics 2020 : GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries,” vol. 71, no. 3, pp. 209–249, 2021, doi: 10.3322/caac.21660.

[3] M. Mori, S. A. Satoko, S. Murasaki, and I. Daniels, “Diagnostic accuracy of contrast-enhanced spectral mammography in comparison to conventional full-field digital mammography in a population of women with dense breasts,” *Breast Cancer*, pp. 1–5, 2016, doi: 10.1007/s12282-016-0681-8.

[4] S. Mittal, “Sampling Approaches for Imbalanced Data Classification Problem in Machine Learning Sampling Approaches for Imbalanced

Data Classification Problem in Machine Learning," *Proc. ICRIC 2019. Springer*, no. January 2020, 2022, doi: 10.1007/978-3-030-29407-6.

[5] T. Biostatistics, *Topics in Biostatistics*, vol. 404. Totowa, New Jersey 07512, 2007. [Online]. Available: https://link.springer.com/book/10.1007/978-1-59745-530-5

[6] C. Dubey, "https://amalaj7.medium.com/logistic-regression-eb2903251107," *Logist. Regres. Made Simple*, 2021.

[7] T. Edition, *Data Mining*, Third Edit. Urbana–Champaign. [Online]. Available: http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf

[8] P. Janardhanan, L. Heena, and F. Sabika, "Effectiveness of Support Vector Machines in Medical Data mining," *J. Commun. Softw. Syst.*, vol. 11, no. 1, pp. 25–30, 2015, doi: 10.24138/jcomss.v11i1.114.

[9] H. Bhavsar and M. H. Panchal, "A Review on Support Vector Machine for Data Classification," *Int. J. Adv. Res. Comput. Eng. Technol.*, vol. 1, no. 10, pp. 185–189, 2012, doi: ISSN 2278 – 1323.

[10] U. Fİdan, E. Uzunhİsarcikli, and İ. Çalikuşu, "Classification of Dermatological Data with Self Organizing Maps and Support Vector Machine Dermatolojik Verilerin Öz Düzenleyici Harita ve Destek Vektör Afyon Kocatepe Üniversitesi Fen ve Mühendislik Bilimleri Dergisi Classification of Dermatological Data," *Afyon Kocatepe Üniversitesi Fen Ve Mühendislik Bilim. Derg.*, no. January, 2020, doi: 10.35414/akufemubid.591816.

[11] M. A. Khan, M. A. K. Khattk, and S. Latif, "Voting Classifier-based Intrusion Detection for IoT Networks," *arXiv e-prints*, no. December, 2021, doi: 10.1007/978-981-16-5559-3.

[12] F. A. B. Kulkarni, Ajay, Deri Chong, *Foundations of data imbalance and solutions for a data democracy*. Elsevier Inc. doi: 10.1016/B978-0-12-818366-3.00005-8.

[13] M. A. Naji, S. El Filali, K. Aarika, E. H. Benlahmar, R. A. Abdelouhahid, and O. Debauche, "Machine Learning Algorithms for Breast Cancer Prediction and Diagnosis," *Procedia Comput. Sci.*, vol. 191, pp. 487–492, 2021, doi: 10.1016/j.procs.2021.07.062.

[14] M. F. Ak, "A comparative analysis of breast cancer detection and diagnosis using data visualization and machine learning applications," *Healthc.*, vol. 8, no. 2, 2020, doi: 10.3390/healthcare8020111.

[15] M. M. Islam, M. R. Haque, H. Iqbal, M. M. Hasan, M. Hasan, and M. N. Kabir, "Breast Cancer Prediction: A Comparative Study Using Machine Learning Techniques," *SN Comput. Sci.*, vol. 1, no. 5, pp. 1–14, 2020, doi: 10.1007/s42979-020-00305-w.

[16] Y. N. Hantoro, "Comparative Study of Breast Cancer Diagnosis using Data Mining Classification," *Int. J. Eng. Res. Technol.*, vol. 9, no. 6, pp. 790–795, 2020.

[17] F. Teixeira, J. L. Z. Montenegro, C. A. Da Costa, and R. Da Rosa Righi, "An analysis of machine learning classifiers in breast cancer diagnosis," *Proc. - 2019 45th Lat. Am. Comput. Conf. CLEI 2019*, 2019, doi: 10.1109/CLEI47609.2019.235094.

[18] Y. Khourdifi and M. Bahaj, "Applying best machine learning algorithms for breast cancer prediction and classification," *2018 Int. Conf. Electron. Control. Optim. Comput. Sci. ICECOCS 2018*, pp. 1–5, 2019, doi: 10.1109/ICECOCS.2018.8610632.

[19] S. B. Sakri, N. B. Abdul Rashid, and Z. Muhammad Zain, "Particle Swarm Optimization Feature Selection for Breast Cancer Recurrence Prediction," *IEEE Access*, vol. 6, pp. 29637–29647, 2018, doi: 10.1109/ACCESS.2018.2843443.

[20] A. Bazila Banu and P. Thirumalaikolundusubramanian, "Comparison of bayes classifiers for breast cancer classification," *Asian Pacific J. Cancer Prev.*, vol. 19, no. 10, pp. 2917–2920, 2018, doi: 10.22034/APJCP.2018.19.10.2917.

[21] V. Chaurasia, S. Pal, and B. B. Tiwari, "Prediction of benign and malignant breast cancer using data mining techniques," *J. Algorithms Comput. Technol.*, vol. 12, no. 2, pp. 119–126, 2018, doi: 10.1177/1748301818756225.

[22] S. Aruna and L. V Nandakishore, "K NOWLEDGE B ASED A NALYSIS OF V ARIOUS S TATISTICAL T OOLS IN D ETECTING B REAST," *Comput. Sci. Inf. Technol.*, pp. 37–45, 2011, doi: 10.5121/csit.2011.1205.

[23] H. Asri, H. Mousannif, H. Al Moatassime, and T. Noel, "Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis," *Procedia Comput. Sci.*, vol. 83, no. Fams, pp. 1064–1069, 2016, doi: 10.1016/j.procs.2016.04.224.

[24] D. W. H. Wolberg, "https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic)," *M.L Repos.*, 1995.

[25] I. Letteri, A. Di Cecco, A. Dyoub, and G. Della Penna, "A Novel Resampling Technique for Imbalanced Dataset Optimization," *arXiv Prepr. arXiv*, pp. 1–23, 2020, [Online]. Available: http://arxiv.org/abs/2012.15231

[26] I. Izonin, R. Tkachenko, N. Shakhovska, B. Ilchyshyn, and K. K. Singh, "A Two-Step Data Normalization Approach for Improving Classification Accuracy in the Medical Diagnosis Domain," *Mathematics*, pp. 1–18, 2022.

[27] S. K. Karlos, Stamatis, Georgios Kostopoulos, "A Soft-Voting Ensemble Based Co-Training Scheme Using Static Selection for Binary Classification Problems," *Algorithms*, 2020, doi: doi.org/10.3390/a13010026.